DOCUMENT RESUME

ED 178 557 TM 009 547

AUTHOR Manning, Winton H.
TITLE Educational Research, Test Validity and Court Decisions.
PUB DATE Apr 77
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, NY, April 4-9, 1977)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Competitive Selection; *Court Litigation; Federal Court Litigation; *Occupational Tests; *Personnel Selection; Predictive Validity; Standardized Tests; Standards; Supreme Court Litigation; Test Bias; Testing Problems; *Test Validity
IDENTIFIERS Construct Validity; Content Validity; *Criterion Validity; Equal Employment Opportunity Commission; Federal Executive Agency

ABSTRACT

A review of court cases, involving the use of various standardized tests, highlights the kinds of test validity and the interpretation by the courts of test fairness in setting personnel selection standards. Guidelines established by the Equal Employment Opportunity Commission (EEOC) are shown to support the use of predictive validity in test-related litigation, while the Federal Executive Agency Guidelines on Employee Selection recognize several other validity models, such as content and construct validity; the acceptance of appropriate measures of training success as criteria in validation studies; a focus on the impact of the totality of the selection process; and an awareness that its guidelines are principles designed to assist people. Court cases discussed include Armstead v Starkville; Tyler v Vickery; U.S. v State of North Carolina; and Washington v Davis. The viewpoint highlighted in this paper supports a broader perspective on test validity models and the need to reconceptualize the theory of test validation so as to balance an emphasis on statistical evidence with the ongoing process of scientific inference. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178557

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

EDUCATIONAL RESEARCH, TEST VALIDITY AND COURT DECISIONS

Winton H. Manning

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Winton H. Manning
ETS
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 009547

Presented at Annual Meeting
American Educational Research Assoc.
April 4-9, 1977 - New York City

EDUCATIONAL RESEARCH, TEST VALIDITY AND COURT DECISIONS

Winton H. Manning

In August 1970 I spent a memorable week in Oxford, Mississippi, where I gave testimony as the only expert witness for the plaintiffs in a federal desegregation case known to some of you as Armstead v. Starkville. The Starkville Municipal School District had finally been compelled to desegregate its dual school system. Subsequent to this there was the familiar pattern of flight of white students into the private academies that then were blossoming, like mushrooms after a spring rain, in the dark earth of the Mississippi Delta. These reduced enrollments in turn led to a reduction in the size of the teaching force, thus affording a tempting opportunity for the school district to get rid of its black teachers, and retain the whites.

I had become involved at the request of the attorneys representing a group of black school teachers who had not been rehired because they had failed to attain the requisite passing scores on the GRE Aptitude and Advanced Tests. This test requirement had been arbitrarily and suddenly introduced by the District on the specious grounds of seeking to upgrade the quality of their teaching staff, but I shared the belief of the minority teachers that the true intent of the District was to make a last ditch

effort to create a "lily white" teaching staff, now that desegregation of their schools was finally coming to pass. It was clear to me that this was a misuse of the GRE Aptitude and Advanced Tests, and for several weeks prior to taking the witness stand in Oxford I had labored away at preparing a lengthy affidavit outlining in impressive detail the multitude of reasons why this use of the tests was inappropriate. Essentially, my whole argument was the classic one of describing all the requirements of a criterion-related test validation effort, and showing how the school district's policy was unprofessional and unsupported by predictive validity studies, or job analyses, as required by recognized standards, including those of APA, AERA, NCME and EEOC.

In addition, I had spent a great deal of time in what must be described as a kind of tutoring of lawyers - concerning the meaning of such things as predictive*, concurrent*, content* and construct validity*, differential prediction, correlation, reliability - statistical significance - in short, the usual kit of psychometric concepts that is familiar to all of you. Like many lawyers I know, these attorneys were impressively fast learners, and they taught me a little law along the way, while they soaked up great gobs of psychometric lingo.

---

*For brief definitions, please refer to the glossary which may be found at the close of the paper.

Although it sounds immodest, I think it would be fair to say that my testimony was a tour de force; in short, it was an exhilarating, emotionally satisfying experience. And, of course, as the lawyers pointed out, it did not hurt the cause of the plaintiff teachers that a Vice President of ETS - the institution responsible for creating the GRE tests - was testifying on their side and saying that the test shouldn't be used for this purpose.

When, a few weeks later, the Court handed down a decision favorable to the plaintiff teachers, I was pleased that our side had won; and my enthusiasm was hardly dampened even when the NEA bureaucracy, in an effusion of anti-test zeal, got things all mixed up. In a colossal obfuscation of the facts NEA managed to convey the impression to the press that I had defended the use of the GRE tests by the Starkville District, rather than having served as an expert witness on behalf of the plaintiff teachers.

Some months later I began to read what others more learned in the law than I, perceived in the Court's opinion in Armstead v. Starkville. What they said caused me to go back and reread the decision of the Trial Court, and then the Appeals Court, and it began to dawn upon me that the Court had gone a lot further than I had ever imagined they would in righting the wrongful policies of the Starkville schools. When the decision finally came down, it was widely interpreted as signifying that,

among other things, it was unlawful for experienced teachers to be required to demonstrate a minimal knowledge of subject matter as measured by standardized tests. Recently I have been publically chastised in the pages of Change magazine (Sherman, 1976) by a writer who is a teacher of mathematics for my shameful role in creating a legal precedent that knowledge of subject matter is irrelevant to the teaching of the subject.

I am here this morning to say, Malcolm Sherman, wherever you are, that you have probably made a very good point. I am a lot older and a little bit wiser than I was in 1970, and if I had to do it over again I'd still mount the witness chair, but I think I would also have tried to say some other things or some things differently that would argue against the Court having drawn a decision so broadly as to suggest that knowledge of subject matter is irrelevant to the performance of experienced teachers. It is not a good social policy for the courts to suggest that the use of academic or intellectual standards in evaluating teachers is a denial of their civil or constitutional rights. As you have said, Mr. Sherman, "...a more profoundly anti-intellectual viewpoint would be difficult to conceive"!

By now you may be asking yourself whether this bit of personal history - the browning of Win Manning, one colleague has called it - has any particular message for educational researchers. I think so, and the message is contained in the remarks to follow.

For several decades psychologists and educational researchers have been telling users of tests that it is necessary to validate interpretations or inferences drawn from tests. Sometimes we have used the language sloppily, implying that it is the test instrument that is somehow validated, but when we speak or write carefully it is generally understood that it is the inferences from the test scores that must be validated - not the test itself, from which, potentially, hundreds of different inferences, in different contexts, might be made.

We have also, quite properly, placed heavy emphasis upon empirical methods of validation - primarily predictive validity designs. Again, somewhat sloppy language has sometimes been used, implying that criterion-related, empirical methods are somehow different from logical or rational approaches to test validation - but again, the professional knows that any effort at validation of inferences involves a complex interplay of empirical observation, rational or logical analysis, in hypothesis formulation and verification. In the past decade these professional views have also found expression in formal statements and in government pronouncements.

It is well known that the EEOC Guidelines, which have been very influential in litigation over the past decade, exhibit a strong bias for predictive (or concurrent) validation designs.

Other approaches, such as content or construct validation, are given short shrift and are "permissible" only if a predictive validity approach is not feasible. This position has more recently come to be regarded by many, if not all, measurement professionals as doubly invidious - in the first sense because it makes little scientific sense to have a general preference for one validation strategy over another and in the second sense because it guarantees a gradual cleavage between an official, bureacratic version of science, and the real article, that is a constantly changing, dynamic corpus of theory, method and knowledge.

The Uniform Guidelines for Employment Selection Procedures promulgated last July, 1976, by the EEOC go a long way toward abandoning the unnecessarily rigid and outdated approach to validation of selection instruments that characterizes the older EEOC Guidelines. The new Uniform Guidelines come down in favor of a more flexible approach that recognizes the advances in psychometric theory, knowledge, and research in the decade since the issuance of the EEOC Guidelines. This new flexibility in the Uniform Guidelines is evident in the acknowledgment of coequality of several validity models; the acceptance of appropriate measures of training success as criteria in validation studies; a focus on the impact of the totality, rather than individual segments, of the selection process, and an awareness that the Uniform Guidelines are principles designed

to assist users, rather than formally applied standards to be rigidly applied in all cases.

Unfortunately, as you may know, the EEOC withdrew its support for the Uniform Guidelines and continues to adhere to its earlier position, thus further complicating the state of affairs with regard to legal requirements for validation in an increasingly politicized environment. It is at least irresponsible and possibly dangerous for the Federal government to inflict upon society and its institutions two sets of guidelines (quasi-regulations, some would say) that are conflicting in their interpretations regarding what is essential in the validation of tests for employment selection.

The EEOC Guidelines of 1970 actually reflect the state of the art as of 1966 and they are increasingly at variance with a gradually deepening understanding of scientifically based validity theory. The Uniform Guidelines of 1976 are better, but far from the best statement that could be made. To some extent, there is also reason to worry a bit about the adequacy of the Standards for Educational and Psychological Tests (1966, 1974) sponsored by APA, AERA and NCME, and upon which the EEOC Guidelines drew heavily as a source. Despite their admittedly beneficial influence, the Standards are also not a finished document, but must be open to change and reconceptualization. It is not my purpose to comment

at length on the Guidelines or the Standards. Perhaps, my views could be encapsulated by recalling to those among you trained in the classics, the judgment of Tacitus on the Emperor Galba -

"Omnium consensu capax imperii nisi imperasset"

which means in plain English:

"Had he never been placed in authority nobody would ever have doubted his capacity for it."

My doubts about the authority of any single-minded approach to test validation that relies almost exclusively upon criterion-related, predictive or concurrent validity designs began sometime during the fallout from Armstead v. Starkville and has continued to grow. More importantly, I think the courts are beginning to have some second thoughts, too, and it is to that I would next like to turn my attention.

I would like to illustrate the changing legal environment by reviewing the outcome of the Tyler v. Vickery case, which dramatizes as well as any the difference between 1970 and today.

In July, 1976, the United States Supreme Court denied plaintiff's petition for review of the 5th circuit's decision in Tyler v. Vickery, thus allowing that decision to stand. This case concerned the use and validation of the Georgia bar examination, which the plaintiffs (aspirants to the bar) claimed was racially biased and invalid. More specifically, it was the plaintiffs'

contention that the undisputed adverse impact of the test, coupled with the lack of a professional test validation study, rendered use of the examination unconstitutional.

A brief summary of the Tyler v. Vickery case is, I believe, revealing with respect to the new disposition of the Court regarding the issue of what constitutes evidence of test validity in instances where an examination does show widely varying performance of minority and majority candidates. In their request for review of the decision of the Court of Appeals, petitioners summarized the evidence as follows:

"In evidence presented in opposition to the motion for summary judgment, petitioners [those denied admission to the bar] showed that the Georgia bar examiners, had for many years, consistently failed a higher proportion of black applicants than white, including failing all forty-one black applicants on the July, 1972, exam; therefore, petitioners argued, the bar examiners had a heavy burden of justification to support the use of a testing procedure that created, intentionally or unintentionally, such a markedly adverse impact on minority applicants. Petitioners argued that the respondents [the Georgia bar examiners] could meet their burden of justification only by reference to scientifically established methods of evaluating the relation of test scores to occupational performance. These scientific testing standards

impose two principal requirements: (1) that a test, such as a bar examination, be demonstrated to be related to job performance, either by correlation of test scores with ratings of actual job performance (criterion-related validity, the preferred method), or by comparison of items and elements of the test with a comprehensive analysis of the elements of the job (content validity), and (2) that the passing score selected has a demonstrable relation to the purpose of the test.

"The testimony offered on the motion for summary judgment showed a sharp dispute on the issues of validity of the test. The bar examiners testified as to their testing procedures, and both petitioners and respondents presented expert witnesses to evaluate those procedures. With respect to the job-relatedness of the test, expert witnesses for both petitioners and respondents agreed that a criterion-related analysis [predictive validation] of the bar examination was feasible, although no such study has been done. (Criterion-related test validation has generally been held, both professionally and judicially, to be the preferred method of showing job relatedness. Douglas v. Hampton, 512 F.2d 976, 986 (D.C. Cir. 1975); Rogers v. International Paper Co., 510 F.2d 1340, 1349 (8th Cir. 1975). Respondents' expert witness testified that a bar examination should have "content validity." [(An examination has "content validity" when "the behaviors demonstrated

in testing constitute a representative sample of behaviors to be exhibited in a job." "An employer cannot justify an employment test on grounds of content validity if he cannot demonstrate that the content universe includes all, or nearly all, important parts of the job." APA Standards, pp. 28, 29.)] Petitioners' expert witness testified that "criterion-related validity" (a more rigorous standard) ought to be demonstrated, and that the record in these cases does not even establish that the bar examination has "content validity." With respect to the passing score, all bar examiners testified that they set the passing score of "70" purely as a matter of judgment. Respondents' expert testified that the judgment alone is not a satisfactory basis for setting a passing score. Petitioners' expert witness added (1) that an empirical study was necessary to establish that the passing score used was not racially discriminatory, and (2) that the passing score adopted by defendants represents a significantly different standard of performance from year to year. Respondents offered no reply to this expert testimony concerning the passing score.

"Petitioners also contended that they had established a prima facie case of purposeful discrimination by the respondent bar examiners by showing, among other factors, that a disproportionate percentage of black applicants had been failed by the bar examiners for many years. Finally, petitioners asserted that they were denied due process of law in that they had no opportunity for

a hearing or review of any kind of the assignment of a failing grade to an applicant. Petitioners presented substantial evidence of arbitrariness in grading that supported the need for a hearing, including variation from year to year of the level of test achievement associated with a grade of "70", with no explanation by the bar examiners for such variations." What was the outcome of this case?

"The majority of the court rejected the principal contention of the petitioners--that the user of a test that produces discriminatory results must prove the job-relatedness of the test--and held that the equal protection clause of the Constitution requires nothing more than that an examination "test skills and knowledge which have a 'logical, apparent relationship' [not a demonstrated one] to those necessary" to the job. 14a, 517 F.2d at 1101. All expert testimony was held to be, as a matter of law, irrelevant to evaluation of the constitutional claim. [(Emphasis supplied.)]

As stated at the outset, the Supreme Court did not grant review to the petitioners, and the decision of the Court of Appeals stands.

My purpose in citing this case is not to draw inferences about other litigation using Tyler v. Vickery as a precedent - that is a job for lawyers to do. Rather, it is to emphasize the

fact that neither the Supreme Court nor the lower courts appear to have been overwhelmed by the heavy emphasis upon predictive validity designs advocated by EEOC; in fact they rejected exclusive reliance on a criterion-related approach to the establishment of job-relatedness. The Court held (and I repeat) that the Constitution requires nothing more than that an examination "test skills and knowledge which have a logical apparent relationship [not a demonstrated one] to those necessary to the job." In denying review of this opinion on appeal, the Supreme Court essentially expressed its agreement with this view.

Let us turn to another case - the U.S. v. State of North Carolina. In this case the issue was whether a state may require a minimum amount of knowledge as a precondition to teacher certification, recognizing, nevertheless, that there is always a substantial gap between the acquisition of a requisite knowledge and effective teaching. (Here the tests are the National Teacher Examinations published by ETS.) The plaintiffs argued forcefully that no test of knowledge of a subject matter covered in the college level teacher training program could be required because there had been "no empirical demonstration that such knowledge is necessary for effective teaching."

In support of this view, the Justice Department employed an expert witness for the plaintiffs, who sought to deprecate the

value of tests of knowledge of subject matter in his effort to nullify the required examinations for teacher certification. Let me quote from his testimony:

> "But the thing, in my opinion, from my teaching experience and what I know about teachers that I have been impressed with, and what I know teachers have to frequently do, and do well, and that is to be able to teach what they don't know, or at least don't know going into teaching the course." Reprinted in Brief Amicus Curiae for Educational Testing Service, U.S. and North Carolina Association of Educators v. State of North Carolina, 1975.

The Court's opinion in this case is instructive, falling much closer to the views expressed in Tyler v. Vickery than to the EEOC Guidelines.

> "We think it is beyond argument that the State of North Carolina has the right to adopt academic requirements and written achievement tests designed and validated to disclose the minimum amount of knowledge necessary to effective teaching...NTE tests do not measure teaching skills but do measure the content of the academic preparation of prospective teachers. We are inclined to think that the NTE tests measure the critical mass of

knowledge in academic subject matter and that a score somewhere on the scale would disclose the knowledge necessary as a prerequisite to effective teaching." (Opinion of the three Judge Court rendered in U.S. et al v. North Carolina, August 27, 1975.)

On balance the Court seems to be saying that reasonable persons know that knowledge is an essential foundation for effective performance, but that good performance depends on a great many other things as well. More important for our discussion today, the Court used the phrase - "we think it is beyond argument..." thus implying that the version of scientific validation deemed essential by the plaintiffs - predictive validity - may be sometimes irrelevant in evaluating the validity of inferences made from tests. Rational judgment is more trustworthy, the Court seems to say.

The third recent case that illustrates the changing mood of the courts is Washington v. Davis.

As most of you know, this case turned on use of an employment test (Test 21) by the District of Columbia Police Department, as one element in the selection of recruits for the Police Academy. In 1970 a class action suit on behalf of all blacks who had failed the test since 1968 was filed in the U.S. District Court, charging that the test had a differential racial impact, and had not been validated according to the accepted

professional standards which means essentially a criterion-related, predictive validity design of the type prescribed by EEOC Guidelines. The District Court upheld the use of the test by the Police Department, but the Appeals Court reversed this decision, holding that the Police Department had not established the test to be "job-related" according to EEOC requirements, and expressing "grave doubts" as to whether validation of an employment test against performance in a "training program" could ever satisfy the applicable legal requirements.

In due course the case was brought to the Supreme Court. Among others, both Division 14 of APA and ETS filed amicus briefs. Both organizations were in essential agreement with respect to the technical questions of test validation but let me summarize the position advocated in the ETS brief.

"Analysis of the requirement of job-relatedness under Title VII of the Civil Rights Act of 1964,...has been confused by the failure to consider separately the two principal issues comprehended by this requirement and by the tendency to utilize technical evidence of test validation alone to measure compliance with the basic legal requirements of fairness and rationality in classifications made for employment selection purposes.

"Courts should consider first the issue of whether the knowledge, skills or other attributes used by an employer to classify prospective employees are fair and rational bases for selection decisions. Employers should be permitted to meet this threshold requirement of job-relatedness by demonstrating that the knowledge, skills or other attributes to be measured in prospective employees are a fair sample of those required in overall job performance or that they are key elements of job performance, the absence of which will create a substantial risk of non-success on the job. If an employer elects to utilize key elements, that procedure should be permitted at each stage of a multi-stage employment process, such as one utilizing a training program, in the same fashion as it is permitted in a single-stage employment process. These judgments are largely independent of any test validation process, and should be made by assessing all the relevant evidence under traditional judicial standards of reasonableness and fairness.

"After determining that a measure of job performance or selected key elements may properly be used as a basis for employment selection, the courts should then turn to the separate question of the fairness and rationality of any standardized tests used. Evidence of test validity may be useful in making these judgments, although there is no single measurement of test validity or

preferred evidence by which the fairness and rationality of the use of a particular test can be assessed under all circumstances. The legal standards for determining the appropriateness of test use should avoid prescription of any one method of test validation and should be stated so as to permit the development of new techniques and appropriate combination of established techniques.

"Such an analysis of the job-relatedness requirement is fully consistent with this Court's decisions and applicable legal and professional standards. Both the District Court and the Court of Appeals in this case failed to separate the two distinct issues involved in the determination of job-relatedness. Both courts over-generalized with respect to the applicable legal standards and failed to consider important evidentiary links in the chain of proof establishing job-relatedness in a multi-stage employment selection process. Further elucidation of the job-relatedness requirement by this Court is both desirable and appropriate on the basis of the record in this case." (Willens, Siemer & Sims, 1975.)

On June 7, 1976, the U. S. Supreme Court, in a landmark decision, ruled that use of Test 21 was permissible, thus reversing the Appeals Court decision. A reading of the opinion of the Court suggests that the Justices hold the following views concerning test validation:

1) It is important to rest decisions regarding the appropriateness of a test used in selection on a rational interpretation of all the evidence--rather than on the frequently narrow criterion-related validity model preferred by EEOC and many lower courts heretofore;

2) There is no serious question about the appropriateness of using assessments of performance in training as a criterion--rather than exclusively demanding "on-the-job-performance" criteria in test validation;

3) The Court rejected the frequently voiced contention that where the test scores show an "adverse impact" for a racial minority, the full burden of justification for continued use of the test falls on the test user who must prove that the test is unbiased and that there is a <u>necessity</u> for its use, even if no racially discriminatory motivation is alleged;

4) It is appropriate to use "key elements," "critical skills," or "minimal competencies" on "significant components" of a job--rather than requiring that a test be validated against the full range of skills, knowledge, and other attributes revealed by a formal job analysis--again, a departure from EEOC Guidelines;

5) It is also appropriate to use a general verbal aptitude test as a selection device when justified on both rational and empirical grounds--rather than restricting ourselves, as do many critics of tests, to more atomistic measures of job competency or performance skills.

It is correct that this decision applies to cases brought as constitutional issues of due process and equal protection, rather than under statutory provisions of the law, but the Court in Washington v. Davis appears to have made clear, in my opinion, what they will say when a case is brought forward exclusively under provisions of Title VII of the Civil Rights Act of 1964.

At this point it may be useful to summarize the import of my comments thus far, for measurement researchers in testing.

The measurement profession's concerns for test validity is of a duration as long standing as the widespread use of standardized tests. The current APA Standards have their roots in formal technical recommendations promulgated by that organization and others nearly twenty-five years ago. Professional literature of the last fifty years reflect extensive inquiries into test validation with a great burgeoning of research inquiry occurring during World War II and in the years following.

In the amicus brief filed for ETS in Washington v. Davis, it was stated that:

"The purpose of the validation work done over the years is to provide information that will assist test developers to construct more useful tests and to assist test users in interpreting test scores, two principal measurement concerns. This validation work has not generally been concerned with developing standards that permit a decision as to whether a test may be used at all, the most relevant legal concern. To some extent, the interpretive assistance found in evidence of test validity assists in making the legal judgment. Such evidence may be necessary to the legal judgment; it should not be substituted for the legal judgment because it was not developed or intended for that purpose.

"Until recently, a number of courts have been developing what appears to be an unrealistic reliance on empirical evidence of test validity. Most of the lower court testing cases thus far have involved the use of aptitude tests. The APA Standards reflect a preference, for measurement purposes, for criterion-related validation of aptitude tests. The EEOC Guidelines have adopted this preference for criterion-related validity as a legal standard for all types of tests. Section 1607.5(a) provides that employers may establish a test to be "job related" through the use of:

'Evidence of content or construct validity, as defined in [the APA Standards],...where criterion-related validity is not feasible.'

"This leap from a preference for criterion-related validity for measurement purposes with respect to some tests to a legal standard with respect to all tests has been reflected in some lower court decisions. Such cases have usually involved aptitude tests and an evident lack of fairness and rationality such that the same result would probably have been reached under any analysis. Other courts, however, have been more sensitive to the need for flexibility in considering validation evidence. Any standard which appears to rely exclusively on statistical evidence of criterion-related validity will be both illogical and unfair to employers who have used standardized tests in a rational and professionally accepted manner.

"Criterion-related validity has inherent limitations as an exclusive basis for legal judgments about test fairness. To establish criterion-related validity it is necessary only to establish a moderately positive correlation between two sets of data for a given group of persons.

"Plainly, for use as a legal standard of fairness and rationality, criterion-related validity should be supplemented by evidence to support at least two ancillary propositions: (1) what

the standardized test actually measured; and (2) what the performance criterion actually measured." (Willens, Siemer, Sims, 1976.)

Recently Harold Gulliksen reminded us of the views of E. F. Lindquist who put it well many years ago when he said:

> "...unfortunately, correlation coefficients have come to be regarded almost solely as 'validity coefficients' ...the correlation between any so-called predictor and any so-called criterion may and often does reveal just as much about the criterion as the predictor [test].... It is the purpose of a [validation study] to report interrelationships among the variables for whatever these relationships may reveal about the student, the high school and the college, as well as about the tests....
>
> An ultimate objective, then, is really to identify and do something abut the <u>other</u> factors...such as, variations in emotional, motivational, and environmental factors affecting the student while in college; in reliability of grades; in appropriateness of instructional materials, procedures and goals with reference to individual student needs; and so on." (Lindquist, 1961.)

Gulliksen goes on to point out that one consequence of the over-emphasis on predictive validity is the corresponding

tendency toward uncritical acceptance of available criteria--school grades, supervisors' ratings, records derived from work samples, and so forth. Available criteria are commonly accepted as received wisdom, and the effort is then made to devise predictor tests that will have as high a correlation as possible with these criteria. When aptitude tests are evaluated it is also "routine procedure to regard high validity coefficients as indicating that the aptitude test is good and should be used, while low validity coefficients indicate that the aptitude test should be discarded or revised. Only rarely is the question raised, 'Should this validity be high or low?'" Gulliksen also appropriately calls our attention to the still relevant wisdom of Alec Wesman, who, in 1950, offered numerous illustrations which demonstrated that a low validity coefficient may be very instructive, enlarging our understanding of the problem, stimulating a critical examination of the criterion, and often leading a careful investigator to revise or discard the criterion rather than to throw out the test. (Wesman, 1950.)

Let me now summarize briefly the points I have tried to make.

1) Like many other psychologists, I had an initial enthusiasm for relying heavily on predictive validation designs as a tool for improving the professionalism of test use, and for bringing about needed social reform--the elimination of irrational impediments to the attainment of equality of opportunity;

2) A decade of experience with the Standards and the EEOC Guidelines suggests that an excess of mindless empiricism is invited by the rigid preference for criterion-related predictive validation that many expert witnesses from the profession of measurement have espoused, and that many courts have accepted;**

3) Major court decisions in the past year indicate that the judiciary is recoiling a bit from what appears to be illogical or irrational conclusions about tests that sometimes follow from purely technical, empirical argument of measurement specialists--especially when doctrinaire adherence to EEOC Guidelines is the crucible in which the test is to be evaluated;

4) There is a need to reconceptualize the theory of test validation so as to strike a more satisfying balance between the emphasis on statistical evidence, and the logical, hypothetical and deductive processes of scientific inference.

In regard to this last point, we are not devoid of voices crying out in the desert of naive empiricism. Wesman (1950),

**A review by the author of sixty employment testing cases decided in federal courts since Griggs reveals a strong preference for criterion-related validity studies, as required by EEOC Guidelines, although by no means are the decisions uniform in this respect.

Lindquist (1961), and Gulliksen (1959) all have warned, decades ago, of the ills that can afflict those who rely too much on predictive validity alone. More recently Messick has proposed a way out of the desert--by reconceptualizing the process of test validation within the framework of construct validity. (Messick, 1975.) In this he acknowledges his debt to Jane Loevinger (1957) who twenty years ago also conceived of construct validity as an overarching conceptual framework, subsuming predictive validity and content validity within its span. Messick argues as follows:

> "Construct validity thus appears to be just as important for educational measurement as for psychological measurement, for there are no adequate options. Content validity does not provide a parallel or alternative evidential basis for interpretation, and specific criterion studies are of sporadic interpretive utility, depending upon particular decision contexts. But construct validity, by linking test behavior to a more general attribute, process, or trait, provides not only an evidential basis for interpreting the processes underlying test scores, but also a rational basis for inferring testable implications of scores from the broader network of the construct's meaning. The usefulness of the test is therefore enhanced for both scientific and applied purposes. Indeed, even the very decision of whether or not to use a test for

> a particular purpose is aided by appeal to construct validity...because many of the possible ramifications of test use are inferable from the test's meaning." (Messick, 1975, pp. 961-962.)

It is not unusual for the course of science to be influenced by the experience accumulated through observing its effects within the social arena. Research on fusion sources of energy, DNA recombination, and saccharin as a carcinogen are all examples within the biological and physical sciences that are currently on the front pages of our newspapers. There is an urgent need for thoughtful reexamination of professional standards of test validation in the light of our experience over the past decade, and in the face of justified impatience and skepticism lately manifested by the courts. Failure to heed these signs will not only bring discredit to the profession of measurement, but more importantly, will deter the useful application of sound measurement practices in the construction of a better and fairer society.

April, 1977.

BRIEF GLOSSARY [1]

Criterion-Related Validities: Predictive and Concurrent Designs

Criterion-related validities apply when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion. Statements of predictive validity indicate the extent to which an individual's future level on the criterion can be predicted from a knowledge of prior test performance; statements of concurrent validity indicate the extent to which the test may be used to estimate an individual's present standing on the criterion. The distinction is important. Predictive validity involves a time interval during which something may happen (e.g., people are trained, or gain experience, or are subjected to some treatment). Concurrent validity reflects only the status quo at a particular time. Under appropriate circumstances, data obtained in a concurrent study may be used to estimate the predictive validity of a test. However, concurrent validity should not be used as a substitute for predictive validity without an appropriate supporting rationale.

---

(1) Taken from: Standards for Educational and Psychological Tests, 1974.

Content Validity

To demonstrate the content validity of a set of test scores, one must show that the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definitions of the performance domain, the users' objectives, and the method of sampling are critical to claims of content validity. An investigation of content validity requires that the test developer or test user specify his objectives and carefully define the performance domain in light of those objectives. The definition should ordinarily specify the results of learning rather than the processes by which learning is either acquired or demonstrated. It should be sufficiently detailed and organized to show the degree to which component tasks make up the total domain.

Construct Validity

A psychological construct is an idea developed or "constructed" as a work of informed, scientific imagination; that is, it is a theoretical idea developed to explain and to organize some aspects of existing knowledge. Terms such as "anxiety," "clerical aptitude," or "reading readiness" refer to such constructs, but the construct is much more than the label; it is a dimension understood or inferred from its network of interrelationships.

BIBLIOGRAPHY

Standards for Educational and Psychological Tests, Washington, D.C.; American Psychological Association, Inc., 1974.

Gulliksen, H. "Intrinsic Validity," American Psychologist, 1959, pp. 511-514.

Gulliksen, H. "When High Validity May Indicate a Faulty Criterion," Research Memorandum 76-10, Princeton, N.J.; Educational Testing Service, October 1976.

Lindquist, E. F. Interpreting the 1961 ACT Research Reports, Iowa City, IO; The American College Testing Program, 1961.

Loevinger, J. "Objective Tests as Instruments of Psychological Theory," Psychological Reports, 1957, Vol. 3, pp. 635-694. (Monograph Supplement No. 3)

Manning, W. H. Some Current Controversies in Educational Measurement, Princeton, N.J., Educational Testing Service, 1976, pp. 1-47.

Messick, S. J. "The Standard Problem: Meaning and Values in Measurement and Evaluation," American Psychologist, Vol. 30, No. 10, October, 1975, pp. 955-966.

Sherman, M. "Anti-Intellectualism and Civil Rights," Change December, 1976, Vol. 8, No. 11, pp. 34-40.

Wesman, A. G. "The Three-Legged Coefficient," Test Service Bulletin Number 40, New York; Psychological Corporation, 1950.

Willens, H., Siemer, D. and Sims, T. Brief Amicus Curiae for Educational Testing Service: Washington v. Davis, October, 1975.